It's the process stupid! Using paradata to explain attrition in the German Internet Panel

Handout for the NR workshop

Peter Lugtig – Utrecht University - and Annelies Blom – University of Mannheim
p.lugtig@uu.nl , blom@uni-mannheim.de

Paradata has quickly become a hot topic in survey practice. In telephone surveys, for example, call sequences give insight in optimal calling patterns, while face-to-face surveys typically use paradata to monitor interviewer performance in the field (Durrant, Maslovskaya & Smith, 2015; Sinibaldi, Durrant & Kreuter, 2014). In web surveys, we are familiar with paradata as a means to monitor respondents' answer behaviour in the form of time stamps (Olson & Parkhurst, 2013).

Paradata are data that are collected during the fieldwork, but are not part of the substantive survey data being collected. They are additional data being reflecting how the survey process envelops. While the collection of paradata is not 'free' (Kreuter, 2013), the production of paradata does not cost much, and in web surveys especially so. Survey methodologists have long been looking for data that can help us understand better how nonresponse and measurement errors occur. Models that rely on socio-demographic variables to detect heterogeneity in nonresponse and measurement error usually have disappointing predictive power. The addition of paradata as covariates into these models may help to improve the explanatory power of such models (Rossmann & Gummer, 2015). This is especially the case in panel surveys, where the same respondents are repeatedly interviewed on the same topics, to detect change. Because nonresponse in a panel survey takes place over a longer period, data about the process of survey interviewing may be particularly helpful to understand the process of nonresponse

The goal of this short paper is twofold: 1. To shed light on why paradata may be useful to understand nonresponse in panel surveys and 2. To show how paradata can be used to understand and predict attrition in the German Internet Panel.

The role of paradata in attrition

Socio-demographic variables can only partly inform us on the leverage-factors that are important to a respondent. **Females** have for example been found to drop out less often than males, but this finding is unlikely to be causal. More likely is that females are generally more inclined to help others and care about society. By consequence 'helping society' is the leverage factor that is 'causing' lower attrition rates for females. Similarly, **living with children** may cause respondents to have less time to complete surveys, while **living with a partner who is also participating** may help to prevent attrition

Rather than modeling attrition by such indirect indicators, paradata may prove a better, direct indicator of the causes for dropout. We can ask respondents directly about how **'relevant'** or **'important'** the survey was, and use that as an indicator for the perceived benefits of survey participation (both in terms of economic and social exchange theory). In the data we will use later in this paper, respondents were for example offered a choice between receiving a reward in cash, vouchers, or donating it to charity. This may indicate to what extent respondents are 'economic' or 'social' respondents, and serve as an indicator of what leverage factors generally are important.

We can ask how **cognitively difficult** the questionnaire was, whether questions were **too personal**, or whether a respondent left **any negative comments, whether they used a mobile device** (this makes it harder to complete the survey studies in this paper) to get an

idea about the costs of survey participation. Similarly, the capabilities of the respondent can be measured by measuring the **median duration** respondents need to complete surveys, and whether a **particular survey took much longer** or shorter than this median. The capabilities can further be studied by looking at the Internet use of the respondent, whether he/she **received a computer** from the panel, or how **old the browser is the respondent** is using

To get an idea on how committed respondents are, we can also use direct indicators from paradata. **The time it takes respondents to complete a survey after the invitation** is sent can serve as an indicator for this, as are **interruptions** or **breakoffs** of the questionnaire.

In short, we believe paradata may serve as more direct indicators of the leverage factors that are important in understanding the respondent's decision to participate. With the exception of the study by Rossmann and Gummer (2015) very few studies have investigated paradata. Rossmann & gummer show that participation history, and survey duration are predictive of later attrition. They did however not explicitly try to link indicators derived from paradata to the survey experience respondents have. Some indicators from paradata may serve well to understand the entire process of survey participation in a longitudinal setting (e.g. median duration, how incentives are spent, how relevant surveys are), while other indicators may serve as short term effects (a negative comment, interruption, or duration of last survey). Paradata may then also help in survey practice to prevent attrition. If we find that particular paradata predict attrition well, we can identify respondents at high risk of attrition, and then intervene to prevent attrition from happening. Lynn (2017) has outlined several ways in which this can potentially be done.

We will now use data from the German Internet Panel (GIP) to illustrate how paradata may be used to understand and predict attrition. For this, we look at attrition in the first twelve waves of the panel., and separately study the predictive power of socio-demographic vaioables and paradata. We also look at whether complex interactions between the socio-demographic variables and paradata, as modelled in Machine Learning may help us to better understand and predict attrition. Our results will show that some but not all types of paradata are predictive of attrition, and that paradata are more powerful predictors than socio-demographic variables.

Methods

Sample

The GIP is a probability-based online panel of the general population in Germany aged 16 to 75. The project was initiated in 2012, when individuals in 4878 households clustered in 250 Primary Sampling Units (PSUs) were randomly sampled by means of an area probability sample.

Each of the selected households received an advance letter containing information about a 15-minute face-to-face interview. This face-to-face recruitment interview was conducted in 2543 households yielding a response rate (AAPOR RR2) of 52.5% at this stage. 2121 of these households contained persons within the target age-range of the GIP, amounting to 3775 age-eligible individuals, who were invited to participate in the online panel. If sampled individuals lived in households without computer and/or internet access they were also invited to participate in the online panel and received the necessary equipment (see Blom et al. forthcoming).

Of these 3775 invited individuals 1578 completed a first welcome interview and thus registered with the GIP. In total, we are using 1483 respondents in our analysis. These are

respondents who completed the core interview, i.e. the first full 20-minute wave of the GIP conducted in September 2012. For more information about the recruitment and initial response in the GIP see Blom et al. (2015).

As respondents are invited every other month for a new wave of the panel survey using e-mail invitations. Reminders to nonrespondents are sent on the second and third Friday after the initial wave-invitation, i.e. weekly after at least one week of fieldwork. Panel members that failed to respond to two consecutive waves are additional reminded by phone on the last week of fieldwork in each month.

Below, we display the most frequent attrition patterns for the first twelve waves of the panel. 43 percent of GIP participants participate in all waves, meaning that 57 percent of panellists miss at least one wave at some point during the course of the panel. The most frequent attrition patterns (2,3 and 5 for example) are respondents who drop out permanently after a specific wave. We also see respondents who miss one wave, but quickly return (patterns 4, 6, 12 and 14), and finally we see respondents who miss a wave, return, and then become nonrespondents again (pattern 7, also within 'other').

Table 1: Fourteen most frequent missing data patterns in first twelve waves of GIP

Pattern	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	Frequency	Percent
1	x	x	x	x	x	x	x	x	x	X	x	x	645	43
2	X	-	-	-	-	-	-	-	-	-	-	-	74	5
3	X	X	-	-	-	-	-	-	-	-	-	-	48	3
4	X	-	X	x	x	x	x	x	x	X	x	x	40	3
5	X	X	x	-	-	-	-	-	-	-	-	-	34	2
5	X	x	x	x	x	x	x	x	x	X	X	-	31	2
6	X	x	x	x	x	x	x	x	x	-	x	x	27	2
7	X	-	x	-	-	-	-	-	-	-	-	-	24	2
8	X	x	x	x	-	-	-	-	-	-	-	-	18	1
9	X	x	x	x	x	x	x	x	x	-	-	-	17	1
10	X	x	x	x	x	x	x	x	x	X	-	-	17	1
11	X	x	x	x	x	x	x	x	x	X	x	-	17	1
12	X	x	x	-	x	x	x	x	x	X	x	x	14	1
13	X	x	x	x	x	-	-	-	-	-	-	-	14	1
14	x	x	x	x	x	-	x	x	x	X	x	x	11	1
other													452	31

Notes: n=1483. X=wave response, - = wave nonresponse. W2: attrition at wave 2. We are only showing attrition patterns with a frequency n >9

Instruments

Table 2 lists the covariates that we are using in this paper to explain attrition. They are closely linked to the reasons for attrition set out earlier in the theory part of this paper, and consist of a set of individual characteristics of the respondent, as well as paradata regarding their panel participation.

In the models that explain the entire process of attrition in the GIP, we face the problem of missing data on the time-varying covariates. We have solved this issue by multiply imputing these values 5 times using the software package MICE (Van Buuren & Groot-Oudshoorn, 2011) in R.3.3.0 (R Core team, 2016).

Table2: predictors used in attrition models

Variables	Scale/coding	Descriptive	Constant, or time-

		statistics	varying
Socio-demographic variables			
Gender	Male=0/female=1	50% female	Constant
Age	13 categories	Median=7, IQR=5	Constant
Education	7 levels	Median=4, IQR=2	Constant
Household Income	15 categories	Median=7, IQR=4	Constant
Employed	No=0, Yes=1	62% employed	Constant
East/West Germany	West=0/East=1, Berlin=2	West= %, east=%, Berlin=%	Constant
Single	No=0, Yes=1	27% = single	Constant
has children	No=0, Yes=1	62% has children	Constant
Single * age	Interaction term	-	Constant
Big 5: openness	Factor score	-	Constant
Big 5: conscientiousness (factor)	Factor score	-	Constant
Big 5: Extraversion (factor)	Factor score	-	Constant
Big 5: Agreeableness (factor)	Factor score	-	Constant
Big 5: neuroticism (factor)	Factor score	-	Constant
Paradata			
Internet experience			Constant
Survey evaluation: interesting	1 (very positive)-5 (very negative)	Mean=2.63	Time-varying
Survey evaluation: relevant	1-5	Mean=2.64	Time-varying
Survey evaluation: different topics	1-5	Mean=2.64	Time-varying
Survey evaluation: too long	1-5	Mean=1.66	Time-varying
Survey evaluation: too difficult	1-5	Mean=1.45	Time-varying
Survey evaluation: too personal	1-5	Mean=2.26	Time-varying
Survey evaluation: general	1-5	Mean=3.56	Time-varying
Whether reminder sent	No=0, Yes,=1	53% sent reminder	Time-varying
Left negative comment at end of questionnaire	No=0, Yes=1	8% left negative comment	Time-varying
Received a PC from panel	No=0, Yes=1	8% received PC from panel	Time-varying
Time between invitation and survey completion	0-29 days	Mean=9.73	Time-varying
How incentives are spent	0= nothing received 1=cash 2=amazon voucher 3=donation to charity	5% = nothing 12%= cash 31% = amazon 52% = charity	Time-varying (waves 5,8 and 11)
Age of browser version	1-100 months	Mean=16.22	Constant
Device used	1=PC/laptop, 2=tablet 3=smartphone	4% smartphone, 2% tablet	
Duration of questionnaire	1-2788 minutes	Median=10.47 min.	Time-varying
Median duration of all questionnaires	1-74 minutes	Median=16.49	Constant
Interruption	No=0, Yes=1	87% = yes	Time-varying
Breakoff	No=0, Yes=1	5% = yes	Time-varying

Notes: descriptive statistics for time-varying covariates shown at values of wave 1.

Results

Do paradata add explanatory power to attrition models? Answer: yes (see table below)

Table 3: T-values of multivariate logistic regression models predicting attrition

	W2	**W3**	**W4**	**W5**	**W6**	**W7**	**W8**	**W9**	**W10**	**W11**	**W12**
Stable predictor											

Intercept	7.74	3.97	3.31	2.73	2.94	3.89	1.46	4.71	.66	2.16	1.10
children	-.82	-1.11	-.22	-1.19	-1.68	-.99	-2.27	-2.05	-2.13	-.10	-1.82
partner	.29	-.38	-.59	.57	-.60	-.01	.29	.31	-.77	.36	1.07
gender	.20	.92	1.51	.80	.47	1.09	.90	.59	1.37	.01	.91
age	.78	.24	.37	.20	.93	1.40	1.88	-.25	1.92	-.33	.64
Education	2.31	2.19	.23	1.11	1.46	.51	1.41	1.43	1.04	.75	2.10
Income	-.80	.18	-.23	-.51	.07	-.22	-.11	.53	-1.13	-.19	-.96
east	1.05	.33	.00	.90	1.31	.84	.66	1.01	.29	.75	.45
medianduration	1.66	-.09	.02	.64	.31	.83	.36	.94	.35	.32	-.13
employed	-1.28	1.22	.06	.78	.59	.91	-.16	.85	1.07	.81	1.16
agreeableness	.46	1.69	.21	-.93	-1.03	-.56	.32	-.58	-.51	.12	.53
openness	-2.11	.75	-.32	-.09	-.65	-1.07	-.28	-.02	.44	-.19	-.88
conscientious	.09	1.38	.47	-.03	-.14	.10	.69	1.17	.78	1.41	.43
neuroticism	-.29	-2.13	-.23	-.92	-.65	.23	.23	.20	-.54	.42	.49
extraversion	-1.30	.70	-.72	-.54	-1.57	-.26	-.10	-.33	-1.37	.38	.52
Internet experience	1.55	2.62	2.48	2.00	1.48	1.76	2.83	1.87	2.49	1.09	3.32
Single	-1.69	1.01	.41	.95	-.25	.17	-.19	-.30	-1.77	-.12	-1.11
Benpc	-1.24	2.58	2.21	-.74	-.37	-.57	.05	-1.46	-.12	-3.11	.19
agebrowser	2.70	.20	.99	-.08	.91	.29	1.15	.66	1.30	-.32	.54
agesquared	-.13	.28	-.14	.23	-.82	-1.19	-1.45	.56	-1.60	.36	.17
Single * age	-.73	.50	.76	.90	1.28	.53	.73	.49	1.24	.63	-.10
Time-varying											
Duration	-.74	-.23	.87	.42	1.23	.33	1.08	-.45	-.64	.10	-.05
smartphone_	-.81	.97	-.84	-1.79	-.57	.11	.25	-.96	.71	.25	-1.25
# days needed to survey	-2.57	-2.12	-1.70	-3.28	-3.03	-2.07	-.80	-3.25	-1.49	-2.23	-1.68
Needed reminder	-3.01	-3.05	-3.74	-1.61	-1.16	-1.49	-1.81	1.38	.21	.85	-.20
Left negative comments	.29	1.23	1.46	3.08	1.14	1.41	2.67	1.14	3.23	3.62	1.89
Multiple sessions	-1.66	-2.40	-.47	-.55	1.61	.16	-.65	1.74	2.06	.77	-.36
Did not complete prev wave	-5.51	-4.46	-5.92	-6.11	-6.87	-6.59	-6.00	-9.59	-9.35	-3.82	-12.54
interesting	-.42	1.78	-.63	-.76	-.33	-1.60	-.75	1.66	.03	.20	2.28
relevant	-.51	2.32	.03	-1.13	1.24	.58	1.38	-.64	-.59	1.18	.93
different topics	-.64	-2.41	1.24	.48	.42	-.32	1.13	-.22	.86	-.73	-.44
too long	-.35	-.12	-2.22	-1.17	-2.06	-.89	-.89	-.20	-.43	-2.80	-1.33
too difficult	-1.34	.49	-.39	.71	.01	1.16	-1.50	-.40	.94	-.10	1.05
too personal	-.09	-.11	.18	-.07	-1.15	-.73	.60	-2.08	-1.92	-1.31	-3.35
general	1.69	-.30	1.65	2.20	-.61	2.10	1.94	-1.15	1.27	.81	2.78
% correctly classified 0-model	79	83	75	76	74	72	71	71	66	64	62
% correctly classified only socio-demographic predictors	79	84	76	77	75	73	73	72	68	66	66
% correctly classified all predictors	88	85	81	82	81	83	82	82	76	75	78
ICC	.43	.40	.40	.40	.40	.46	.48	.43	.44	.49	.45
% correctly classified all predictors Multilevel model	95	94	92	94	94	95	95	95	95	96	95

Note: n=1483. T-values are based on the pooled Wald-statistics of the predictors in each of 11 regression models predicting attrition (0) or completion (1) of at each wave (2 to 12).
ICC: Intraclass Correlation Coefficient

Do Machine learning models models help?

Answer: only for understanding which paradata patterns will lead to attrition (decision trees shown at workshop)